Assessing Forecast Quality of HII Flood Forecast Service in Chao Phraya River Basin

Kay Khaing Kyaw[1,*] Theerapol Charosesuk[2] Watin Thanathanphon [3] and Piyamarn Sisomphon[4]

[1,2,3,4] Hydro-informatics Modelling Section, Hydro-Informatics Institute, Bangkok, THAILAND

**Corresponding author; E-mail address: kay@hii.or.th*

Abstract

Hydrometeorological forecasts are essential to water management plans including early warning and flood damage prevention. Forecasting models have varying levels of skill depending on the forecast location and period of the year. Measure of skills can have a strong influence on how forecasts impact decisions related to water management, and they must be communicated to the users of the forecasts. Various forecast verification methods are available for assessing the multiple facets of forecast performance including notions such as accuracy, reliability and sharpness. This paper describes a variety of complementary performance metrics to verify Hydro Informatic Institute (HII)'s flood forecasts in Chao Phraya River Basin. The accuracy of the forecasts is evaluated using the continuous rank probability score (CRPS) which quantifies the difference between a forecast distribution and observation. The sharpness of forecasts is calculated using the ratio of inter quantile range (IQRs) of streamflow forecasts and a historical reference. The reliability of forecasts is also considered using attribute diagrams and Kolmogorov-Smirnov (KS) test. In addition, this paper applies the traditional continuous verification methods and statistics such as Bias, Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Nash-Sutcliffe Efficiency Coefficient (NSE), Coefficient of determination (R^2) and Pearson Correlation Coefficient (r). The comparison of the forecast and observed discharge indicate that the MIKE11 model can predict well. The trends are similar in almost all key stations and the overall correlation is acceptable. This study definitely answers the question regarding the correlation between the forecast and observed streamflow and the performance of the forecasts. Based on the verification statistics, it was demonstrated that HII's flood forecasts are reliable.

Keywords: Flood forecasting, Forecast verification, Forecast accuracy, Forecast sharpness, Forecast reliability

1. Introduction

The term forecast means a prediction of the future state and the forecast verification means the process of accessing the quality of a forecast. The forecast is compared or verified against a corresponding observation of what actually occurred, or some good estimate of the true outcome. The verification can be qualitative or quantitative. In either case, it should give information about the nature of the forecast errors. Forecast quality is required to monitor the accuracy and improvement of the forecasts over time. Forecast quality is not the same as forecast value. A forecast has high quality if it predicts the observed conditions well according to some objectives or subjective criteria. It has value if it helps the user to make a better decision. Hydro-informatics Institute Thailand (HII) has a flood forecasting system in the Chao Phraya River Basin and this study will use the flood forecasting (discharge data) from that system to evaluate the performance of the forecasts.

2. Study area and Data

The study area covers the Chao Phraya River Basin and the key stations are selected in five rivers; Ping river, Wang river, Yom river, Nan river and Pasak river.

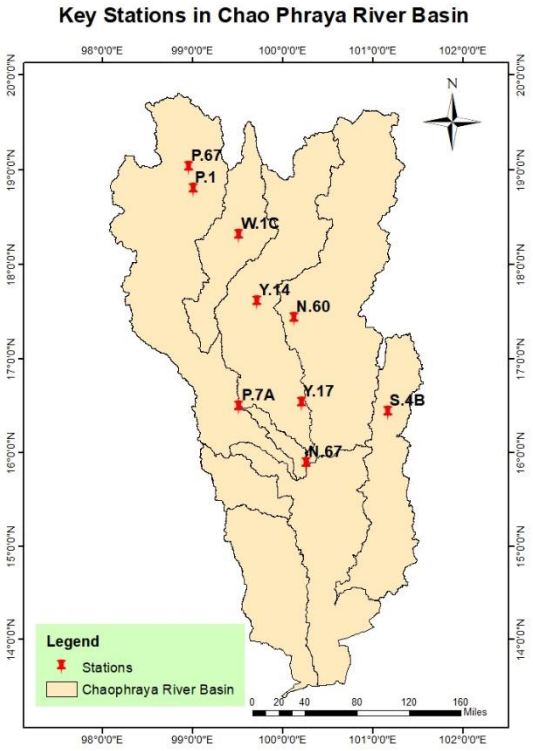

Fig. 1 Location of key stations in Chao Phraya River Basin

Key stations are selected based on having regular long-term maintenance to generate representative series of river system and locations free of backwater effects from any existing, ongoing or future. The generated datasets; observed and forecasted discharge from HII's flood forecasting system are collected and used in this study.

Table 1 Observed and Forecasted Discharge from Key Stations

Stations	River	Frequency	Period
P.67	Ping River	Daily	2017-2020
P.1	Ping River	Daily	2017-2020
P.7A	Ping River	Daily	2017-2020
W.1C	Wang River	Daily	2017-2020
Y.14	Yom River	Daily	2017-2020
Y.17	Yom River	Daily	2017-2020
N.60	Nam River	Daily	2017-2020
N.67	Nam River	Daily	2017-2020
S.4B	Pasak River	Daily	2017-2020

3. Methodology

3.1 Forecast Accuracy

Forecast Accuracy is the level of agreement between the forecast and the observation. The difference between the forecast and the observation is the error. The lower the errors, the greater the accuracy. The continuous rank probability score (CRPS) metric quantifies the difference between a forecast distribution and observation as follows (Hersbach,2000);

$$CRPS = \frac{1}{N} \times \sum_{i=1}^{N} \int_{-\infty}^{\infty} [F_i(y) - H_i\{y \geq y_0\}]^2 \, dy \quad (1)$$

where F_i is the cumulative distribution function (cdf) of the forecast of the year i, y is the forecast variable (here discharge) and y_0 is the corresponding observed value. $H_i\{y \geq y_0\}$ is the Heaviside step function that equals to 1 when the forecast values are greater than the observed values and equal to 0 otherwise. The CRPS summarizes the reliability, sharpness and bias attributes of the forecast. The perfect forecast, i.e. a point forecast that matches the actual value of the predicted quantity, has CRPS=0.

Linear error in probability (LEPS) measures the error in probability space as opposed to measurement space, where $CDF_0(F_i)$ and $CDF_0(O_i)$ are cumulative distribution function of forecasts and observations (Zhang & Casey, 2000). The range of LEPS is between 0 and 1 and the perfect score is 0.

$$LEPS = \frac{1}{N} \sum_{i=1}^{N} |CDF_0(F_i) - CDF_0(O_i)| \quad (2)$$

3.2 Forecast Reliability

Forecast reliability is related to the correspondence between the distribution of forecasts and the distribution of observations. It is critically important so that water managers can confidently eliminate least plausible options in their water allocation and delivery planning which invariably involves extensive scenario analysis [4]. The reliability diagram, also called the attribute diagram, plots the observed frequency against the forecast probability. This can show how well their observed frequencies match to the expected event probabilities [5]. Reliability is indicated by the proximity of the plotted curve to the diagonal. The deviation from the diagonal gives the conditional bias. If the curve lies below the line, this indicates over forecasting and points lying above the line indicate under forecasting.

In statistics, the Kolmogorov-Smirnov (KS) test is a non-parametric test of the equality of continuous, one dimensional probability distributions that can be used to compare a sample with a reference probability distribution (one sample KS test) or to compare two samples KS test. The p value from KS test is used to categorize the reliability rating which is derived form p

value obtained form KS test at significance thresholds varying from 1% to 10%. A p value greater than 10 % implies a strong degree of confidence in forecast reliability, while a p value less than 1% implies a poor probability of accurate forecast distributions [2]. Also, Kolmogorov's D statistics enables to test whether the empirical distribution of data is different than a reference distribution.

3.3 Forecast Sharpness

The sharpness of forecasts is evaluated using the ratio of inter-quantile ranges (IQRs) of discharge forecasts and a historical reference. The following definition is used.

$$IQR_q = \frac{1}{N}\sum_{i=1}^{N} \frac{F_i(100-q)-F_i(q)}{C_i(100-q)-C_i(q)} \times 100\% \quad (3)$$

where IQR_q is the IQR value corresponding to percentiles q, $F_i(q)$ and $C_i(q)$ are respectively the q^{th} percentiles of forecast and the historical reference for year i. An IQR_q of 100% indicates a forecast with the same sharpness as the reference, an IQR_q below 100% indicates forecasts that are sharper than the reference, and an IQR above 100% indicates forecasts that are less sharp than the reference (Woldemeskel et al., 2018). In this study, IQR_{99} ; the IQR at the 99^{th} percentile is used in order to detect forecasts with unreasonably long tails in their predictive distributions.

3.4 Commonly used performance metrics

Standard verification methods are also used in this study to evaluate the forecast performance.

3.4.1 Mean Error

Mean Error can give the average forecast error. The range is between -∞ to ∞ and the perfect score is zero.

$$Mean\ Error = \frac{1}{N}\sum_{i=1}^{n} (F_i - O_i) \quad (4)$$

whereas, *N* = number of samples, F_i = forecast discharge, O_i = observed discharge.

3.4.2 Bias

Bias is the correspondence between the mean forecast and mean observations.

$$BIAS = \frac{\frac{1}{N}\sum_{i=1}^{N} F_i}{\frac{1}{N}\sum_{i=1}^{N} O_i} \quad (5)$$

3.4.3 Mean Absolute Error (MAE)

MAE is the average magnitude of the forecast errors. The range is from 0 to ∞ and the perfect score is 0.

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |F_i - O_i| \quad (6)$$

3.4.4 Root Mean Square Error (RMSE)

RMSE is error computed from summation of mean square of observed and computed values. It can be considered as average value of error.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (F_i - O_i)^2} \quad (7)$$

3.4.5 Nash Sutcliffe efficiency coefficient (NSE)

NSE is frequently used to quantify the accuracy of hydrological predictions. It can answer a question as to how well the forecast predicts the observed time series. The range is between -∞ and 1 and the perfect score is 1.

$$NSE = \frac{\sum_{i=1}^{n} (F_i - O_i)^2}{\sum_{i=1}^{n} (O_i - \bar{O})^2} \quad (8)$$

3.4.6 Coefficient of determination (R^2)

The coefficient of determination (R^2) is a measure that assesses the ability of a model to predict or explain an outcome in the linear regression setting.

$$R^2 = \frac{\sum_{i=1}^{n}[(x_i - \bar{x})(y_i - \bar{y})]^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (9)$$

where x_i and y_i are observed and forecasted data and $\bar{x}$ and $\bar{y}$ are averaged data of them.

3.4.7 Pearson correlation coefficient (r)

The Pearson coefficient is a type of correlation coefficient that represents the relationship between two variables that are measured on the same interval or ratio scale. The Pearson coefficient is a measure of the strength of the association between two continuous variables.

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}} \quad (10)$$

where x_i and y_i are observed and forecasted data and $\bar{x}$ and $\bar{y}$ are averaged data of them.

4. Results and Discussion

4.1 Fitting probability distribution

Before analyzing the data, it is necessary to adjust the probability distributions to the observed and the forecasted (simulated) discharge. This study used three types of goodness

of fit tests, such as the Kolmogorov- statistics, the Crkamer-Mises statistics and the Anderson- statistics. The best probability distribution is then chosen for both datasets at each key station. For example, the statistics at Y.14 station (Yom River), are shown in the following tables. Weibull distribution is the best fit for Y.14 station, and the parameters of shape and scale are determined according to the test. After that, the cumulative distribution functions are calculated using those parameters.

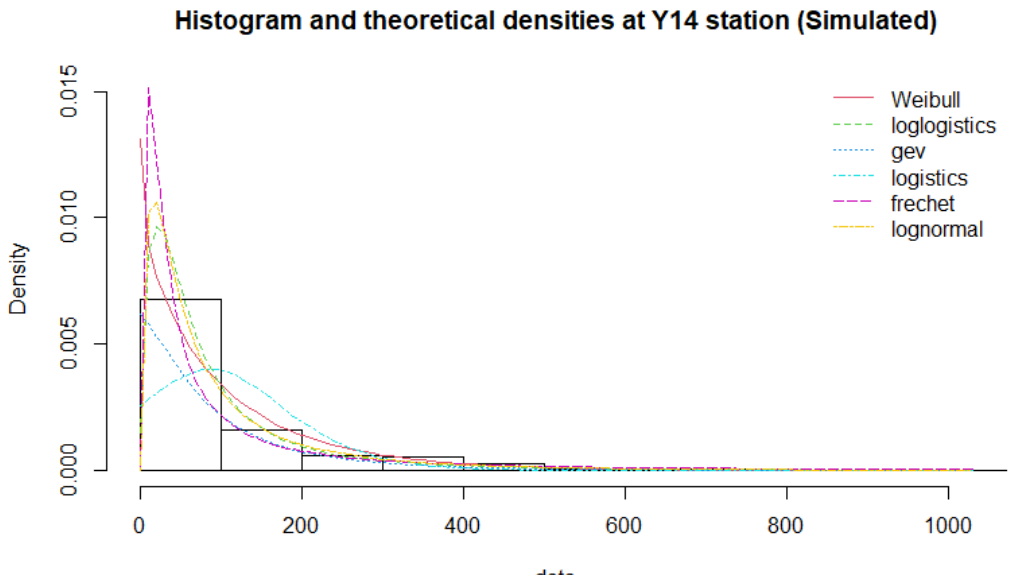

Fig.2 Histogram and theoretical densities at Y.14 station (simulated).

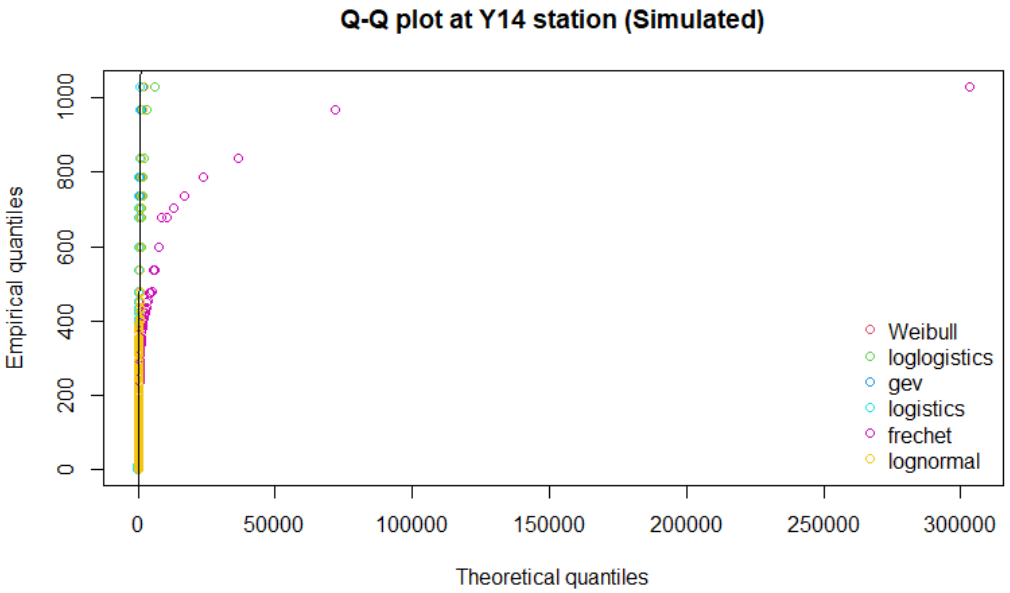

Fig.3 Q-Q plot at Y.14 station (simulated).

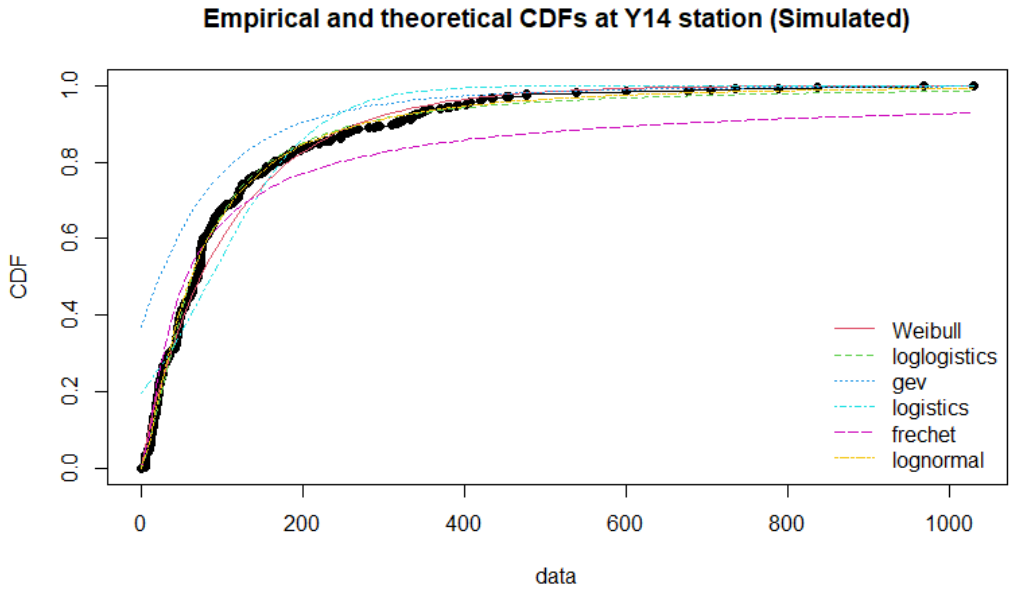

Fig.4 Empirical and theoretical CDFs at Y.14 station (simulated).

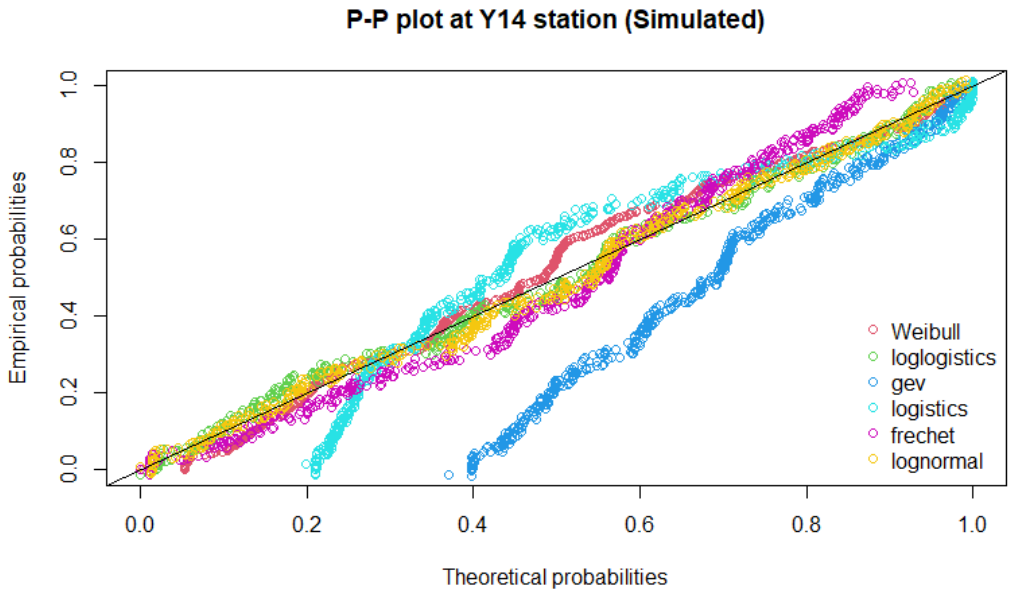

Fig.5 P-P plot at Y.14 station (simulated).

Table 2 Statistics Result for Simulated Discharge at Y.14 Station.

Goodness-of-fit statistics	Weibull	Log normal	Gamma	Log logistics	GEV	Logistics	Burr	Frechet
Kolmogorov-Smirnov statistic	0.087	0.054	0.103	0.048	0.395	0.209	0.380	0.109
Cramer-von Mises statistic	0.741	0.204	1.057	0.204	23.69	4.159	24.84	1.664
Anderson-Darling statistic	4.621	1.154	5.783	1.439	112.8	30.491	120.7	11.314

Table 3 Statistics Result for Observed Discharge at Y.14 Station.

Goodness-of-fit statistics	Weibull	Log normal	Gamma	Log logistics	GEV	Logistics	Burr	Frechet
Kolmogorov-Smirnov statistic	0.081	0.134	0.090	0.123	0.386	0.220	0.347	0.152
Cramer-von Mises statistic	0.842	1.134	0.993	1.304	21.53	3.994	17.91	3.006
Anderson-Darling statistic	5.363	6.939	6.244	6.244	106.5	29.379	89.29	22.343

4.2 Continuous ranked probability score and Linear error in probability score

Both scores are within the range between 0 and 0.5 at all main stations so forecast accuracy is decent and appropriate. LEPS and CRPS scores for all key stations are shown in the following figures.

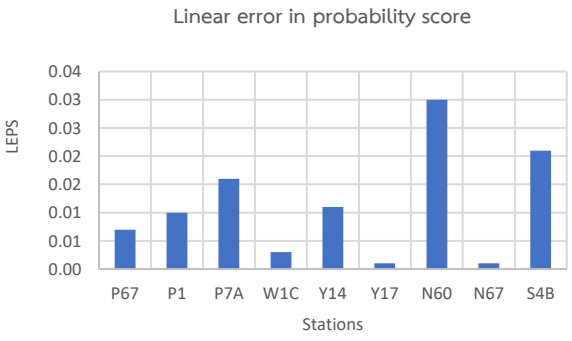

Fig. 6 Linear error in probability score

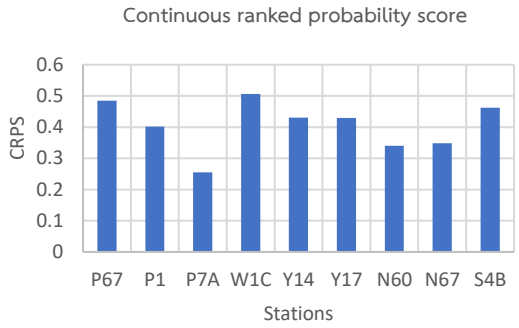

Fig. 7 Continuous ranked probability score

4.3 Forecast reliability test results

Reliability of forecasts is evaluated in key stations using reliability plots and KS tests. Reliability plots represent the curve that can say the observed frequency for a certain probability, and KS test D statistics show the difference between the predicted and observed discharge empirical cumulative dispersion functions. P values from KS test may also provide the response that the data is or is not reliable. For example, the P value for P.67 station is 0.053 so that the forecast is considered to be reliable since the P value is rated as reliable by more than or equal to 5 percent. The figure 8 and 9 show the reliability plots and KS test diagram at P.67 station.

4.4 Forecast Sharpness test results

The Interquartile Ranges (IQRs) for all main Chao Phraya Basin stations are calculated for forecast sharpness. The forecast is less sharp than the observed discharge for P.7A, P.67, N.67, and W.1C stations have predictions sharper than the comparison observed. The forecasts of remaining Stations are as accurate as those of reference. The results can be seen clearly in the bar chart below.

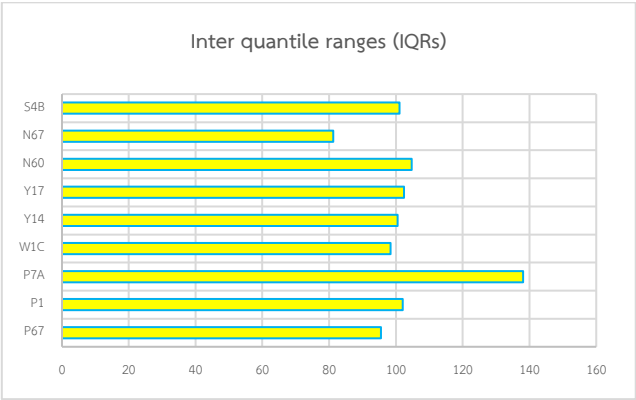

Fig. 10 Inter quantile ranges (IQRs) for key stations

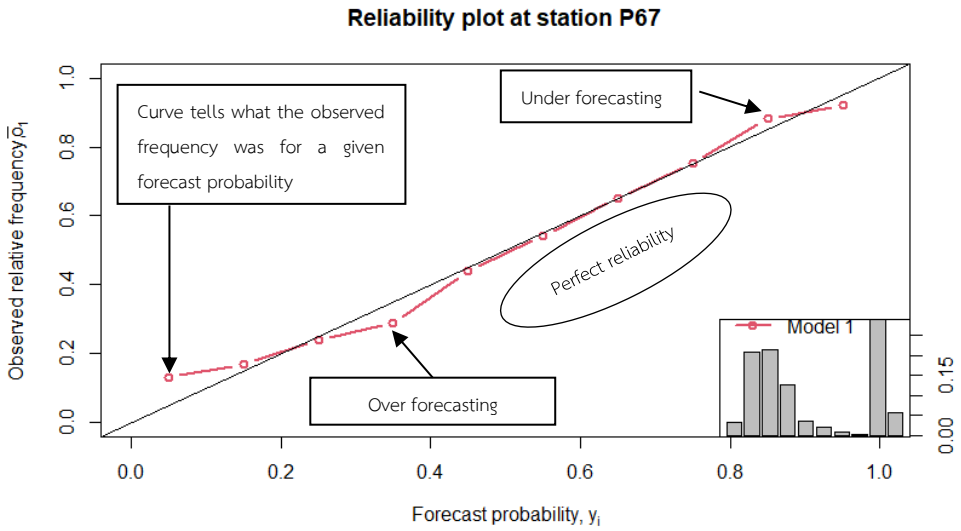

Fig.8 Reliability plot at P.67 station

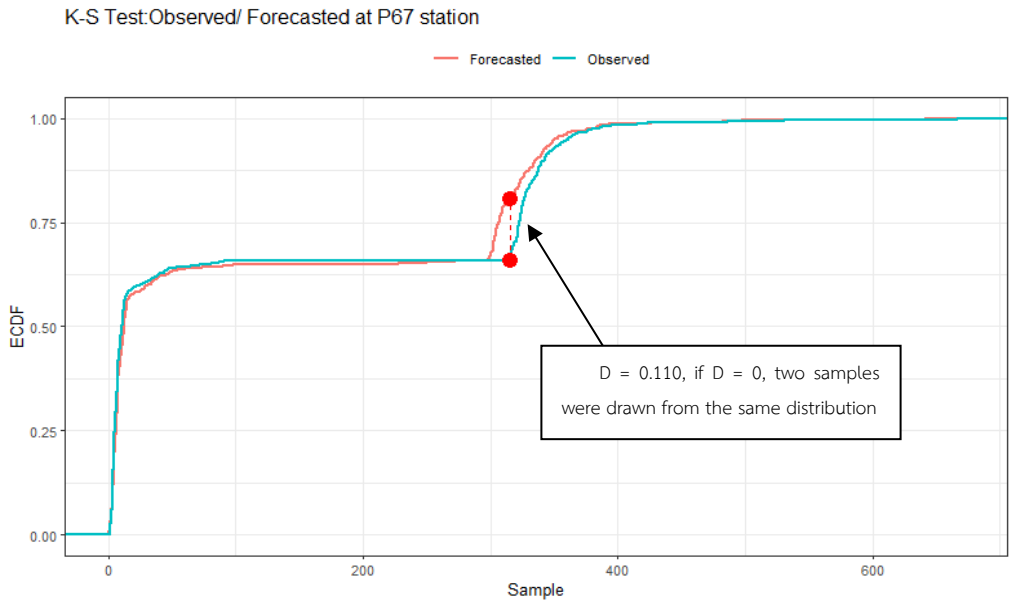

Fig.9 KS test at P67 station.

4.5 Standard verification results

The MAE and RMSE values for S.4B, P.67, W.1C, Y.14 and Y.17 are lower than others. Although the remaining stations have high values, they are also appropriate since the values of RMSE and MAE are considered to be less than half the standard deviation of measured results, and therefore either is suitable for model evaluation. The result of NSE, R^2 and r shows that most stations are almost 1 and therefore well-connected with the observed and predicted discharges.

There are some reasons of using these standard verification metrics. MAE and RMSE measures average error, weighted according to the square of the error. It does not indicate the direction of deviations. The RMSE puts greater influence on large errors than small errors, which may be a good thing if large errors are especially undesirable, but also encourage conservative forecasting. NSE and R^2 are frequently used to quantify the accuracy if hydrological predictions, whereas correlation coefficient measures how do the points of a scatter plot are to

a straight line and it does not take forecast bias into account. Therefore, all of kinds of parameters are need to be considered in evaluating the performance of forecasts.

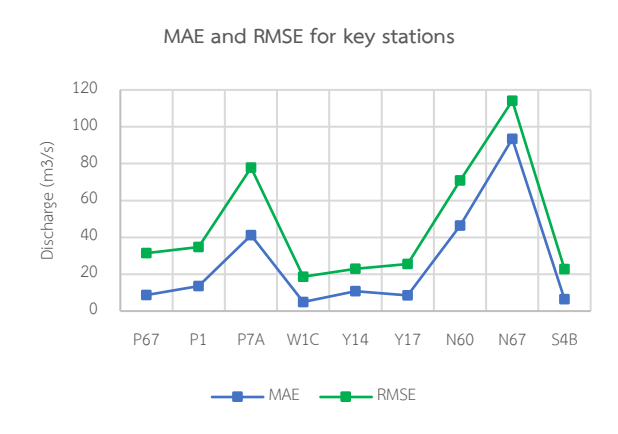

Fig. 11 MAE and RMSE values for all key stations

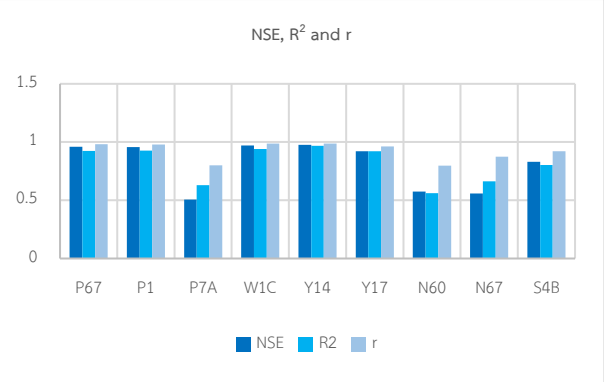

Fig. 12 Correlation coefficients for all key stations

Table 4 Nash Sutcliffe efficiency coefficient (NSE) for key stations

NSE ranges	Station Names	Number of stations	Remarks
0.9 <NSE <1	P.67, P.1, W.1C, Y.14, Y.17, S.4B	6	Very good model performance
0.8 <NSE <0.9	P.7A, N.60, N.67	3	Good model performance

Table 5 Coefficient of determination (R^2) for key stations

R^2 ranges	Station Names	Number of stations	Remarks
0.9 < R^2 < 1	P.67, P.1, W.1C, Y.14, Y.17	5	Very Strong Correlation
0.8 < R^2 <0.9	S.4B	1	Strong Correlation
0.6 < R^2 < 0.8	P.7A, N.60, N.67	3	Moderate Correlation

Table 6 Pearson correlation coefficient (r) for key stations

R ranges	Station Names	Number of stations	Remarks
0.9 < r < 1	P.67, P.1, W.1C, Y.14, Y.17, S.4B	6	Very good correlation
0.8 < r < 0.9	P.7A, N.60, N.67	3	Good correlation

4.6 Taylor diagram

Taylor diagrams [3] provide a way of graphically summarizing how closely a pattern matches observations. The similarity between two patterns is quantified in terms of their correlation, their centered root means square difference and the amplitude of their variations. Figure 13 is a Taylor diagram which shows how it can be used to summarize the relative skill. Statistics for key stations were computed and colored dots were assigned to each station considered. The position of each dot appearing on the plot quantifies how closely that station's forecast pattern matches observations. Consider station P.7A, the green dot, for example. It's pattern correlation with observations is about 0.8. The two contours lie with label 100 and 200 indicate the RMS values and it can be seen that in the case of station P7A, the centered RMS error is about 100 m^3/sec. The standard deviation of the simulated pattern is proportional to the radial distance from the origin. For station P.7A, the standard deviation of the simulated field (about 129 m3/sec) is clearly greater than the observed standard deviation (110 m3/sec).

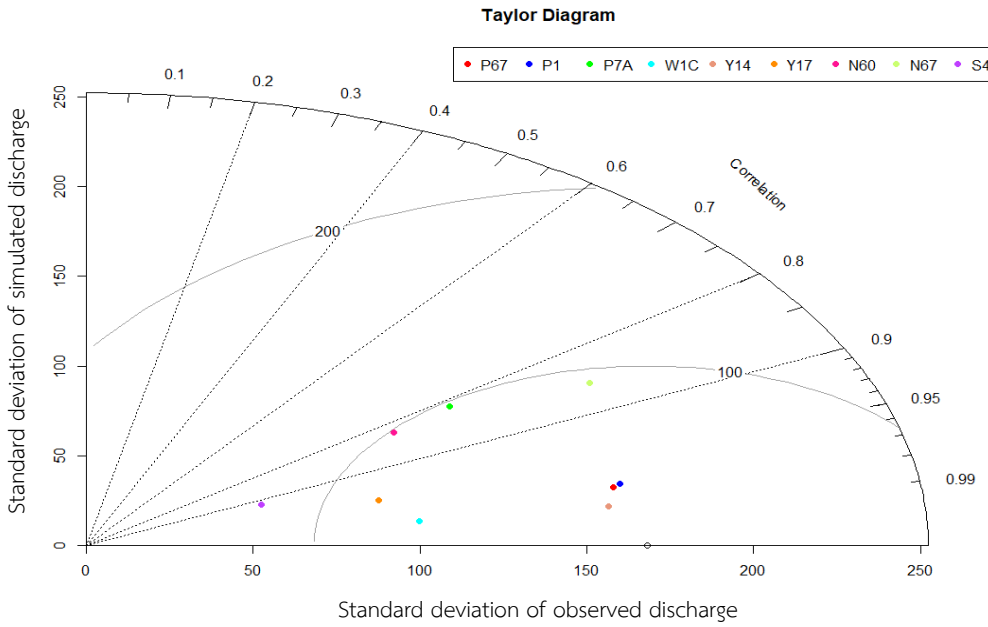

Fig.13 Taylor Diagram

5. Conclusions

This study focused on evaluating the performance forecast of the operational flood forecasting service in the Chao Phraya River Basin. For water managers and decision makers to use these predictions, it should be accurate and have a better sharpness than climatology. We investigated streamflow forecasts evaluating schemes based on verification metrics employing CRPS and LEPS for forecast accuracy, KS test and attribute diagram for forecast reliability, IQRs for forecast sharpness and other standard verification methods for forecasts of continuous variables, namely MAE, BIAS, RMSE, R^2, NSE and r. The analytical findings are obtained as follows.

1. Continuous ranking probability scores and linear error in the probability scores of most stations are closer to zero so that the key stations are in high accuracy.
2. There are three stations (N.67, P.67, W.1C) that have forecasts sharper than the reference so that they can be assumed as the best stations in terms of sharpness.
3. The p values can show the reliability of the data according to the KS test. Out of 9 stations, 4 stations have a p value of more than 5 percent, making them highly reliable. 2 stations are moderately reliable and the remaining stations are unlikely to have a reliable forecast.
4. When considering MAE and RMSE together, only two stations have values greater than half the standard deviation and the remaining stations can be considered to be appropriate for model evaluation.
5. The NSE, R^2 and r values can show the correlation between forecasts and observations. Five stations fall in the range above 0.8 and four stations fall in the range under 0.8. Overall, it can be seen that there is a strong correlation between forecasts and observed discharge at all stations.
6. The Taylor diagram shows that most stations have Root Mean Square (RMS) errors below 100 m^3/s and fall within a range of correlations from 0.8 to 0.99. The simulated patterns of the stations Y.14, N.60 and P.1 are in good agreement with the observations. These stations have relatively high correlation and low RMS errors.

After taking all the verification metrics into account, two stations in Ping River (P.67, P.1) and one station in Wang River (W.1C) are the best stations in forecasting flood. The rest of the stations do more or less well in each verification test. In this study, many kinds of verification metrics are used to evaluate the forecasts because each method has characteristics that reflect forecast behavior. Those results are necessary for decision making for earlier flood alerts, when the weighting of different components

of the flood forecasting system need to be optimized. Therefore, it can be concluded that HII flood forecast service in Chao Phraya River Basin performs very well in terms of forecast reliability, sharpness and accuracy.

Acknowledgement

Data for this study are provided by the Hydro-Informatics Institute (Thailand). Authors thank the anonymous reviewers for constructive comments and feedback that helped us substantially improve the paper.

References

[1] Hersbach, H. (2000). Decomposition of the continuous ranked probability score for ensemble prediction systems. Weather and Forecasting, 15(5), 559–570.

[2] Tuteja, N. K., Zhou, S., Lerat, J.,Wang, Q. J., Shin, D., and Robertson,D. E.: Overview of Communication Strategies for Uncertaintyin Hydrological Forecasting in Australia, in: Handbook of Hydrometeorological Ensemble Forecasting, edited by: Duan,Q., Pappenberger, F., Thielen, J., Wood, A., Cloke, H. L., and Schaake, J. C., Springer Berlin Heidelberg, Berlin, Heidelberg,1–19, 2016.

[3] Taylor, K. E. (2001). Summarizing multiple aspects of model performance in a single diagram. Journal of Geophysical Research: Atmospheres, 106(D7), 7183-7192

[4] Zhang, H., & Casey, T. (2000). Verification of categorical probability forecasts. *Weather and Forecasting*, *15*(1), 80–89.

[5] Turco, M., & Milelli, M. (2007). Towards Operational Probabilistic Precipitation Forecast. Group, (9), 56–62.

[6] Woldemeskel, F., McInerney, D., Lerat, J., Thyer, M., Kavetski, D., Shin, D., … Kuczera, G. (2018). Evaluating post-processing approaches for monthly and seasonal streamflow forecasts. Hydrology and Earth System Sciences, 22(12), 6257–6278.